Elections and Campaign Finance in State Legislatures

Selina Wu

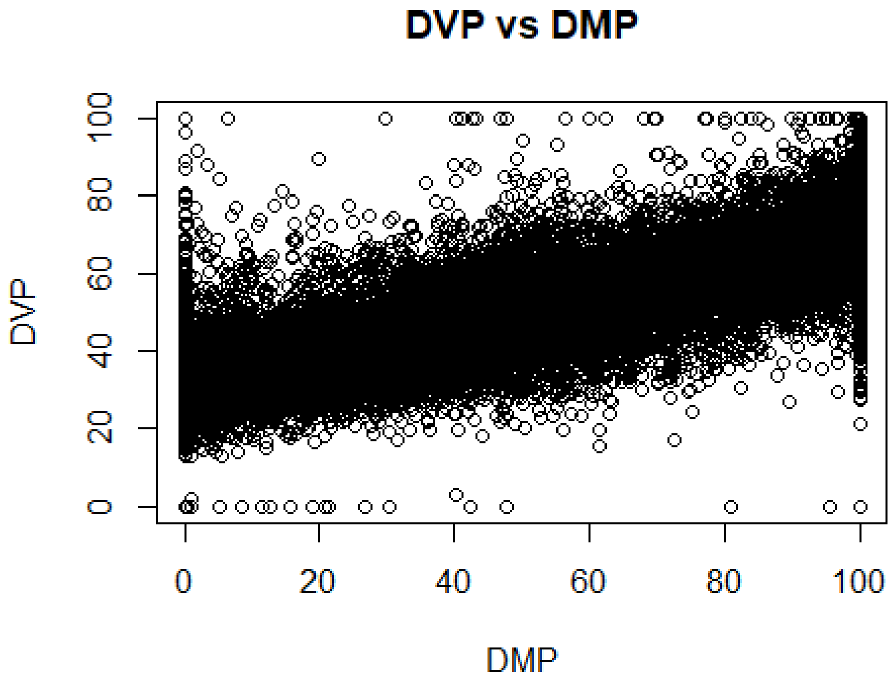
March 4, 2020

```r
setwd("C:/Users/selin/OneDrive/Documents/Data
Projects/PoliSci/2_CampaignFinance")

cf <- read.csv("campaign_finance.csv")
```

Democratic vote percentage and Democratic percentage of all campaign contributions

```r
plot(cf$dem_money_pct, cf$dem_vote_pct, xlab="DMP", ylab="DVP", main="DVP vs
DMP")
```



The Democratic vote percentage (DVP) appears to increase as the overall Democratic contribution percentage (DMP) increases. The two variables seem to have a positive linear relationship, as an increase in one variable seems to be accompanied by an increase in the other variable.

Average DVP and DMP by state

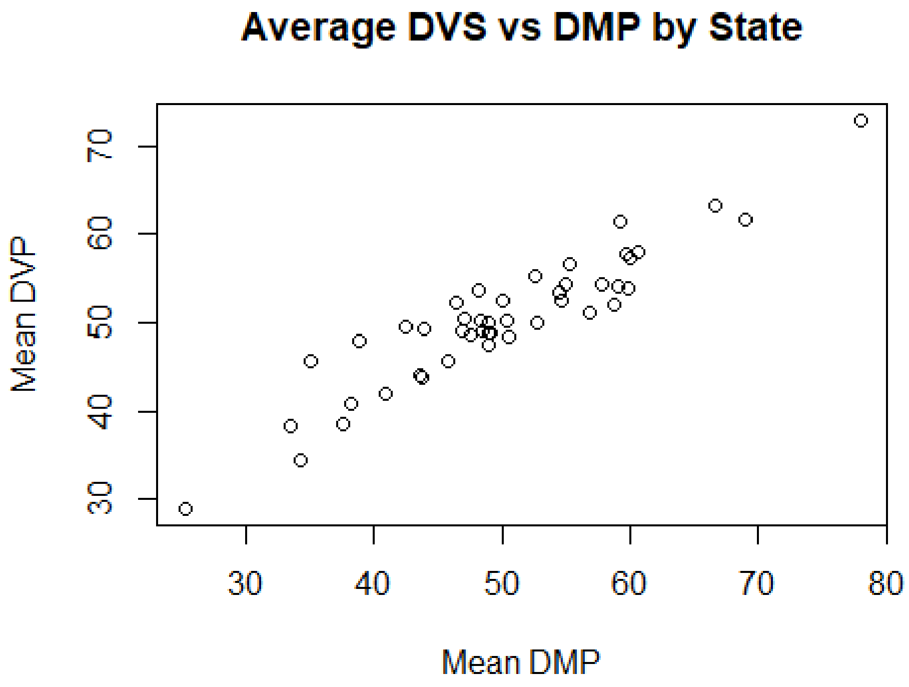
In order to compute the average DVP and DMP for each state, the aggregate function is used to find the mean values by state.

```
state_avgs <- aggregate(cf$dem_vote_pct, list(cf$state), mean, na.rm=T)
state_avgs$DMP <- aggregate(cf$dem_money_pct , list(cf$state), mean, na.rm=T)

colnames(state_avgs) <- c("State", "Mean_DVP", "DMP")
colnames(state_avgs$DMP) <- c("State","Mean_DMP")
state_avgs$DMP$State <- NULL
```

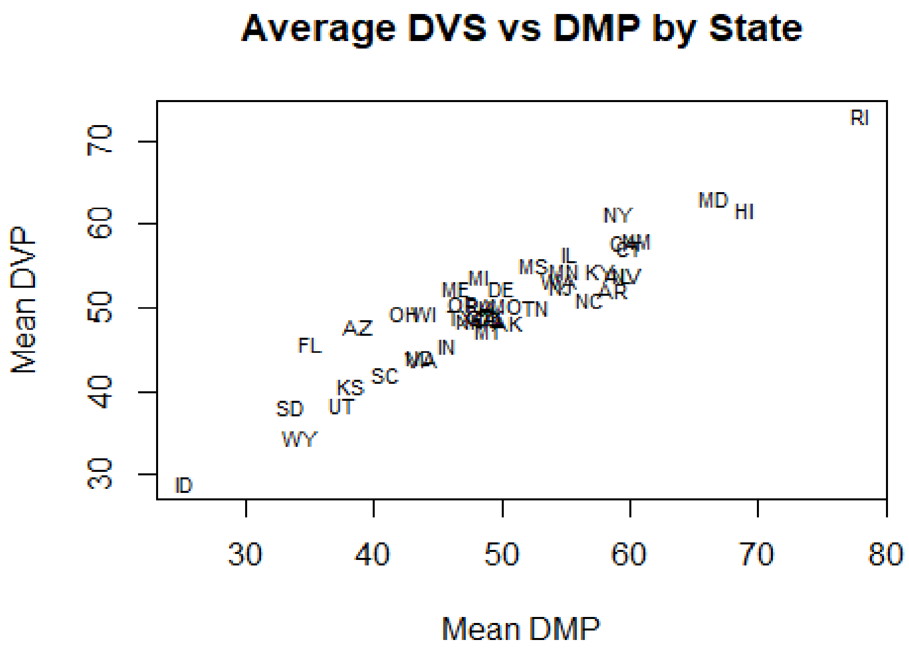
From the data frame state_avgs, the two variables are plotted in a scatterplot.

```
plot(state_avgs$DMP$Mean_DMP, state_avgs$Mean_DVP, xlab="Mean DMP",
ylab="Mean DVP", main="Average DVS vs DMP by State")
```



Using state abbreviations:

```
plot(state_avgs$DMP$Mean_DMP, state_avgs$Mean_DVP, col="white",
       xlab="Mean DMP", ylab="Mean DVP", main="Average DVS vs DMP by State")
text(state_avgs$DMP$Mean_DMP, state_avgs$Mean_DVP, state_avgs$State,
cex=0.70)
```



As the average Democratic money percentage increases for each state, the average Democratic vote percentage also appears to increase. The relationship between Democratic vote share and Democratic money percentage appears to take on a positive linear relationship. For example, in Idaho, the average Democratic money percentage was low, and so was the Democratic vote share. On the other hand, Rhode Island on average received a lot of Democratic funding, and had an average Democratic vote percentage that was very high. T

How well do contributions predict election outcomes?

The difference between the Democratic vote share (DVS) and Democratic group percentage (DGP) is calculated through DVP-DGP. The average difference is -0.6019207%.

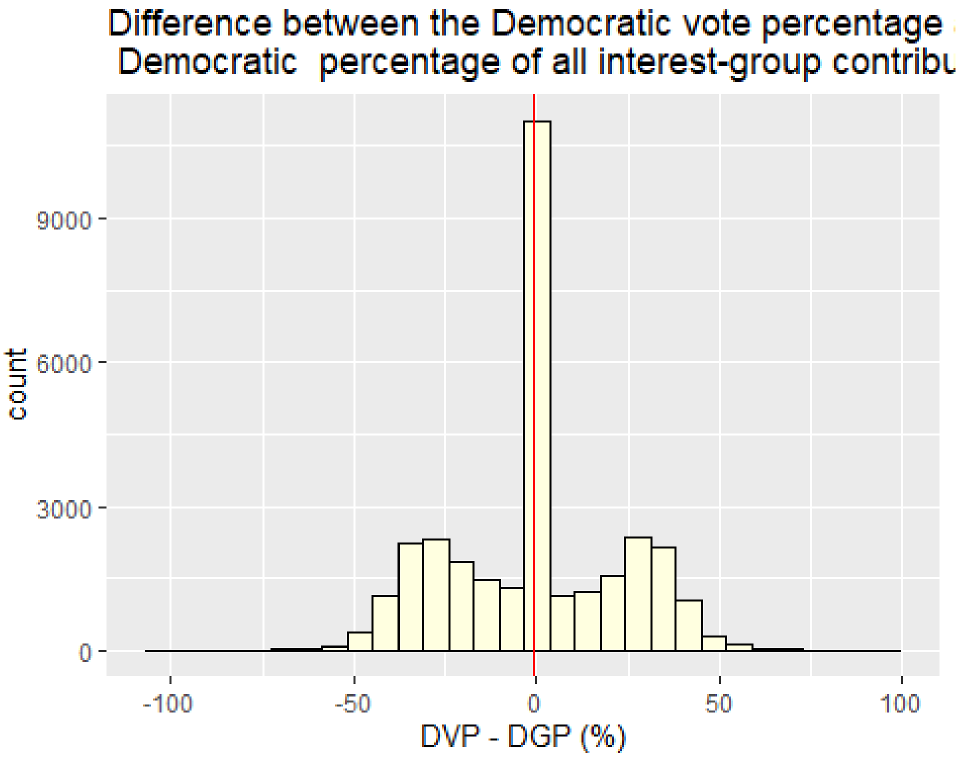
```
dvp_DGP <- cf$dem_vote_pct - cf$dem_group_pct
cf$dvp_DGP <- dvp_DGP
meanVPGP <- mean(cf$dvp_DGP)

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

ggplot(cf, aes(x=dvp_DGP))+geom_histogram(color="black", fill="light
yellow")+
  labs(title="Difference between the Democratic vote percentage and the \n
Democratic  percentage of all interest-group contributions", x= "DVP - DGP
(%)")+
  geom_vline(xintercept = meanVPGP, colour="red")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

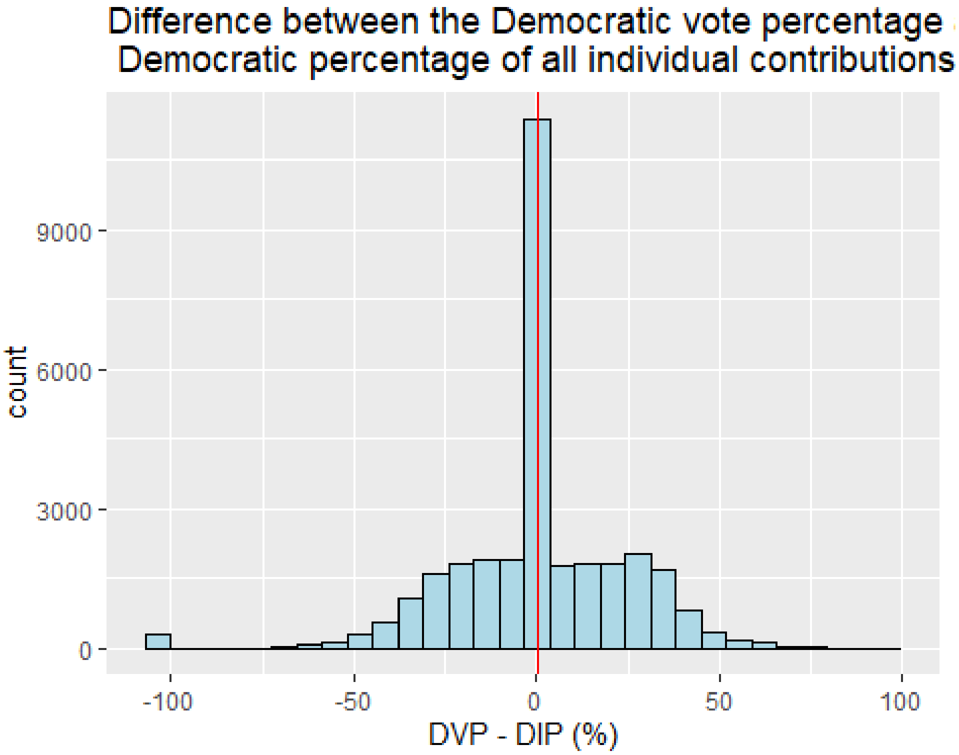


The difference between the Democratic vote share (DVS) and Democratic individual percentage (DIP) is calculated through DVP-DIP. The average difference is 0.625356%.

```
dvp_DIP <- cf$dem_vote_pct - cf$dem_indiv_pct
cf$dvp_DIP <- dvp_DIP
meanVPIP <- mean(cf$dvp_DIP)

ggplot(cf, aes(x=dvp_DIP))+geom_histogram(color="black", fill="light blue")+
  labs(title="Difference between the Democratic vote percentage and the \n
Democratic percentage of all individual contributions", x= "DVP - DIP (%)")+
  geom_vline(xintercept = meanVPIP, colour="red")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histograms show that the DGP and DIP are fair predictors of the DVP, as the average errors between the DGP/DIP and the actual DVP are extremely close to 0. The histogram also shows that the frequency of the differences between DVP and DGP/DIP being 0 is extremely high compared to when there is a more notable difference. If the model is not preditive, the histograms would be more uniform, as the differences between the contributions and the actual vote percentage would be more widespread rather than close to 0.

Since the average difference between the prediction (DGP or DIP) and the actual value (DVP) is close to zero, the percentages of the two kinds of contributions can be considered unbiased predictors of the Democratic vote percentage.

Though both variables appear to be good predictors since each of their average difference is close to 0 for each model, I would say the DIP might be a better predictor of DVP, as the histogram shows that as the errors get larger, the number of elections with showing those errors greatly decreases. On the other hand, there are seemingly more elections with higher error values when the differences are around 25-40% for the histogram with DGP as a predictor.

Prediction: Democratic win anytime Democratic contribution percentage exceeds 50.

###How often does campaign contribution inaccurately predict the outcome of the election?

Although a small number of elections had group or individual percentages equaling 50%, a vote share percentage over 50% is be considered a win, while that of under 50% is considered a loss.
A binary variable, dem.Win, is created to indicate if the Democratic candidate wins the election. Two more binary variables are created – one for the group contributions and one for individual contributions – indicating when these predictions are 'wrong.'

```
dem.Win <-ifelse(cf$dem_vote_pct >50, 1, 0)

dGp.pred.W <- ifelse(cf$dem_group_pct > 50, 1, 0)

dIp.pred.W <- ifelse(cf$dem_indiv_pct > 50, 1, 0)
```

The percent of cases in which the prediction is opposite of real election outcomes is calculated for the group and individual contributions as follows.

```
dGp.Wrong <- (sum(dGp.pred.W != dem.Win)/length(dem.Win) * 100)
dGp.Wrong

## [1] 9.893948

dIp.Wrong <- (sum(dIp.pred.W != dem.Win)/length(dem.Win) * 100)
dIp.Wrong

## [1] 19.25292
```

It appears the group donor predictions are wrong approximately 9.89% of cases, which is less often than the individual donor predictions wrongly predicting approximately 19.25% of the cases.

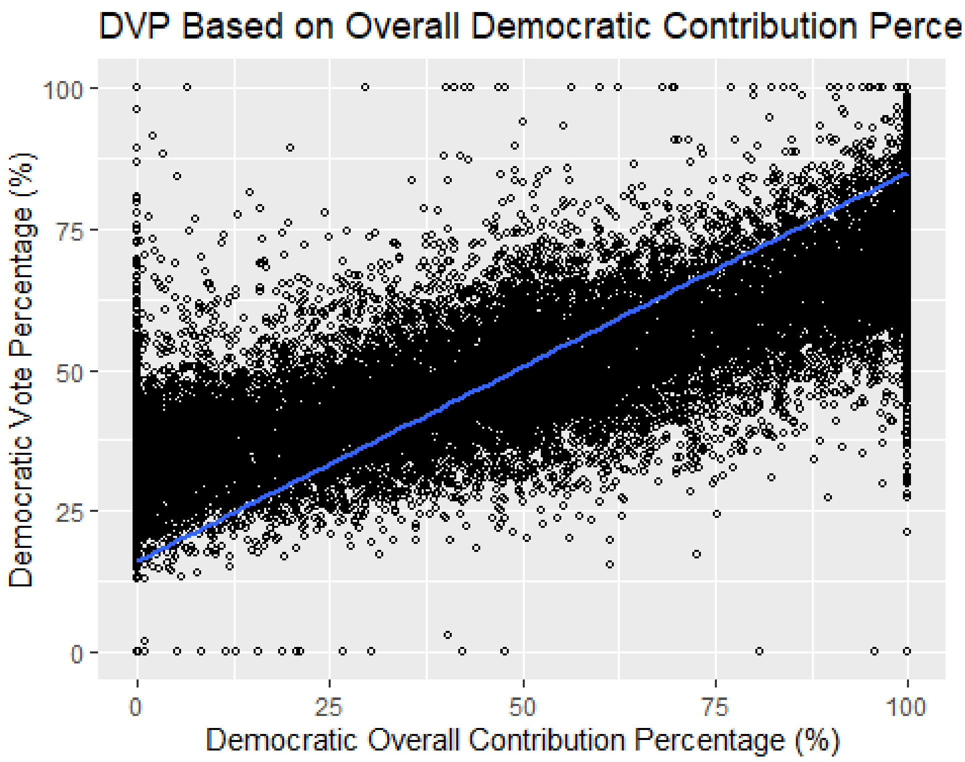
Regression predicting DVP from DMP

```
reg <- lm(dem_vote_pct~dem_money_pct, data=cf)
reg

##
## Call:
## lm(formula = dem_vote_pct ~ dem_money_pct, data = cf)
##
## Coefficients:
##   (Intercept)  dem_money_pct
##        16.102          0.689

ggplot(cf, aes(x=dem_money_pct, y=dem_vote_pct)) + geom_point(shape=1,
size=1) +
  xlab("Democratic Overall Contribution Percentage (%)") +
  ylab("Democratic Vote Percentage (%)") +
  ggtitle("DVP Based on Overall Democratic Contribution Percentage") +
  stat_smooth(se=F, method="lm")

## `geom_smooth()` using formula 'y ~ x'
```



The regression predicts the Democratic vote percentage based on the overall Democratic contribution percentage. The regression has a positive slope of 0.689, indicating that for each 1% increase in Democratic overall contribution, the Democratic vote percentage is predicted to increase by 0.689%. When the Democratic money percentage is at 0%, the predicted Democratic vote percentage is only at 16.102%.

```
pred <- predict(reg)
residuals <- reg$residuals

sum_SR <- sum(reg$residuals^2) #sum of squared residuals
rmSE <- sqrt((1/nrow(cf))*sum_SR)
rmSE

## [1] 14.79055
```

The residuals (difference between the predicted and actual DVP value) are calculated. The root mean square error is the average prediction error of the regression, so this value was calculated using the formula for RMSE.

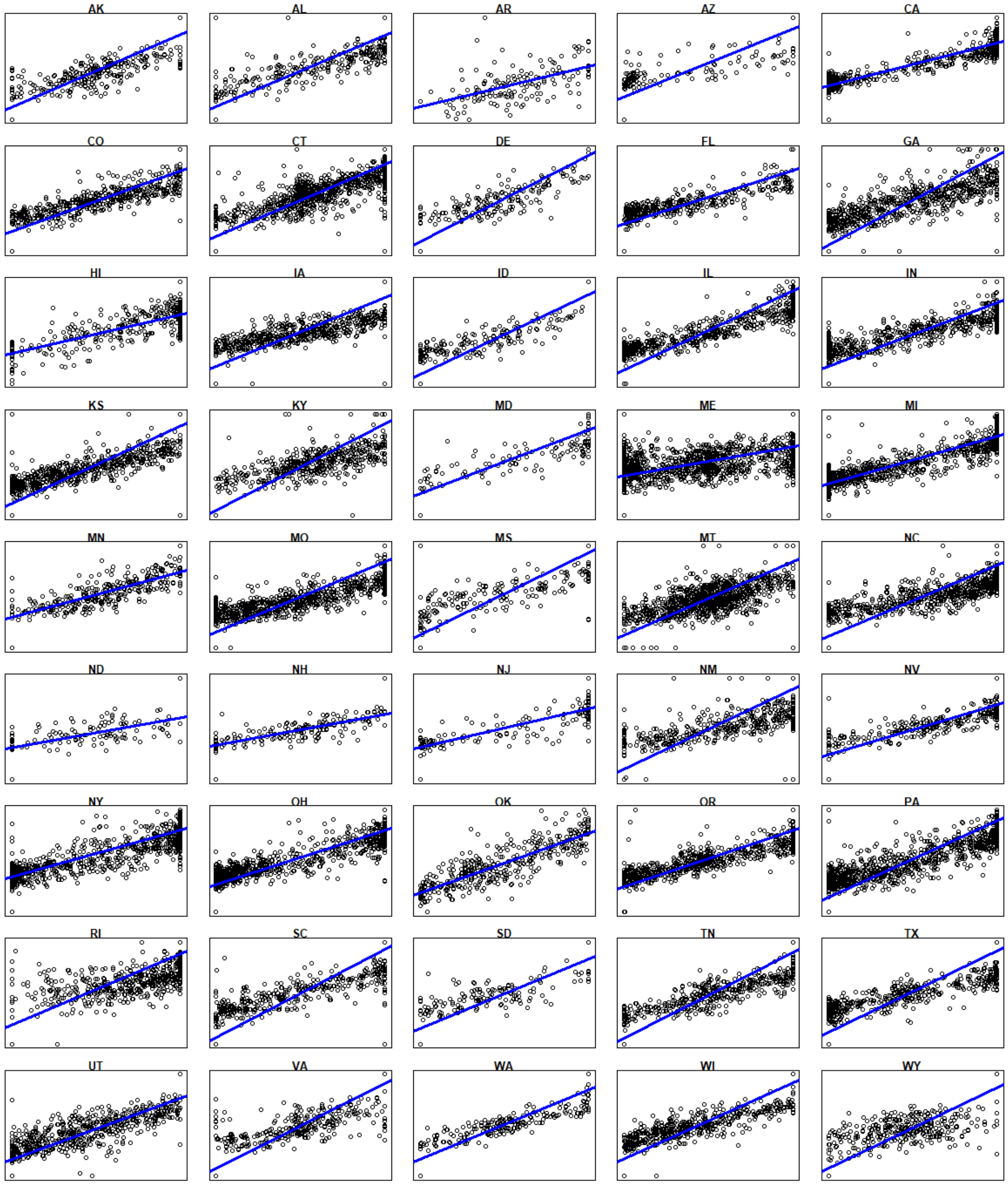
The average prediction error is 14.791%, meaning that based off this regression model, the predicted Democratic vote percentage was higher or lower by, on average, 14.791% than what it actually was. This average error is not close to zero, and the plot supports this by showing a handful of data points lying too high or low from the regression line. This error value can also be due to there being so much data. There were so many elections with varied results, a lot of which do not align closely with the prediction model. With this data, it is difficult to produce a regression model with very accurate predictions.

DVP vs DMP by state

```
par(mfrow=c(9,5), mar=c(1,1,1,1), oma=c(3,3,3,3))

for (i in state_avgs$State){
  plot(cf$dem_money_pct[cf$state==i], cf$dem_vote_pct[cf$state==i],
       xlab="DMP", ylab="DVP", xaxt="n", yaxt="n", main=i)

  temp <- cf[cf$state==i,]
  reg.state <- lm(dem_vote_pct ~ dem_money_pct, data=temp)
  abline(reg.state, col="blue", lwd=3)
}
```



There appears to be a consistent link between Democratic money percentage and Democratic vote percentage. For all of the states, as the DMP increases, the DVP seemingly increases as well. This is further supported by the regression lines with positive slopes.